BIOLOGIA PLANTARUM 54 (1): 54-60, 2010

Comparison of three genetic similarity coefficients based on dominant markers from predominantly self-pollinating species

A. BEHARAV[1]*, M. MARAS[2], M. KITNER[3], J. ŠUŠTAR-VOZLIČ[2], G.L. SUN[4], I. DOLEŽALOVÁ[3], A. LEBEDA[3] and V. MEGLIČ[2]

Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel[1]
Agricultural Institute of Slovenia, Hacquetova 17, SI-1000 Ljubljana, Slovenia[2]
Faculty of Science, Palacký University in Olomouc, CZ-78371 Olomouc-Holice, Czech Republic[3]
Biology Department, Saint Mary's University, Halifax, NS, B3H 3C3, Canada[4]

Abstract

Three genetic similarity coefficients were estimated and compared for their usefulness: simple matching (S_{SM}), Jaccard's (S_J) and Dice's (S_D), all based on dominant markers data from individuals representing predominantly self-pollinating species. AFLP markers were used to analyze 139 *Phaseolus vulgaris* L. (common bean) and 67 *Lactuca saligna* L. (least lettuce) accessions, and RAPD markers were used to analyze 110 *Triticum dicoccoides* Koern. (wild emmer wheat) accessions. Similar discriminating structure and power based on the three genetic similarity coefficients was found for each of the three species. This discriminating power was high for both *P. vulgaris* and *L. saligna* but moderate for *T. dicoccoides*. With closely related individuals, as in our study, the absence of a band in two individuals should be due to an identical cause inherited from the same ancestor. Accordingly we propose the use of S_{SM}, which alone out of the three examined coefficients involved shared absence of DNA bands, as contributing to genetic similarity. When RAPDs are employed, inferences about population structure and nucleotide divergence should be made with prudence as the nature of genetic variation uncovered by RAPDs is often unclear.

Additional key words: closely related individuals, *Lactuca saligna*, negative matches, *Phaseolus vulgaris*, *Triticum dicoccoides*.

Introduction

Characters coded by two or a few states are especially suitable for computation of association coefficients. In the most common model, association coefficients are computed with two-state characters, which for convenience are coded 0 and 1. When character states are compared over pairs of columns in a conventional data matrix, the outcome can be summarized in a conventional 2 × 2 frequency table (Sneath and Sokal 1973). We can sum the cells of each diagonal of the 2 × 2 table to obtain two frequencies, one of matches and the other of mismatches. This is desirable when the meaning of "positive" and "negative" can vary from the presence to the absence of two alternative states of a binary character.

The introduction of DNA-based marker techniques allows direct comparison of different genetic material independent of environmental influences (Weising *et al.* 1995). Molecular markers have been commonly used in recent decades for various applications (Cordeiro *et al.* 2008, Reif *et al.* 2005), where the proper choice of a similarity coefficient S, or of a dissimilarity coefficient $D = 1 - S$ (following the terminology of Gower 1985), between operational taxonomic units (OTUs) is important and depends on various factors. Reif *et al.* (2005) examined 10 similarity coefficients widely used in germplasm surveys, with special focus on applications in plant breeding and seed banks. They suggested the term "allelic informative" if allele frequencies can be determined from the molecular marker data, and the term

Received 14 November 2008, *accepted* 25 April 2009.
Abbreviations: AFLP - amplified fragment length polymorphism; CCC - cophenetic correlations coefficients; *D* - dissimilarity; OUT - operational taxonomic unit; RAPD - random amplified polymorphic DNA; *S* - similarity; UPGMA - unweighted pair group method with the arithmetic averages.
Acknowledgements: The molecular work for this research was supported by grant MSM 6198959215 (Ministry of Education, Youth and Sports of the Czech Republic); in part by grant J4-9476-0401-06 from the Slovenian Research Agency; and by a postdoctoral grant from the Israeli Council of Higher Education to G.L. Sun.
* Corresponding author; fax (+972) 4 8246554, e-mail: beharav@research.haifa.ac.il

"allelic non-informative" if they cannot. They suggested that when the marker data are "allelic non-informative" the estimates of coefficients *D* between OTUs under consideration can be calculated by one of three coefficients, based on the absence or presence of observed bands or signals described by Sneath and Sokal (1973) as follows: Dice's similarity coefficient (S_D) (Dice 1945): $Sij = 2a/(2a+b+c)$, which is equivalent to the Nei-Li coefficient (Nei and Li 1979); Jaccard's similarity coefficient (S_J) (Jaccard 1908): $Sij = a/(a+b+c)$; and Simple matching similarity coefficient (S_{SM}) (Sokal and Michener 1958): $Sij = (a+d)/(a+b+c+d)$, where *Sij* is the similarity between two OTU individuals, *i* and *j*; *a* = bands shared by both individuals; *b* = bands present in *i* but not in *j*; *c* = bands present in *j* but not in *i*; and *d* = number of bands absent from both individuals. In contrast to S_{SM}, both S_J and S_D do not involve shared absence of DNA bands.

Materials and methods

Thirty-six randomly screened recent studies (1998 - 2008: reference list is available upon request) quantified the degree of similarity among individual OTUs by one out of two allelic non-informative markers, namely random amplified polymorphic DNA (RAPDs) or amplified fragment length polymorphism (AFLPs). Altogether they used 44 similarity coefficients as follows: 25 (56.8 %) the S_J coefficient, 13 (29.6 %) the S_D coefficient, and 6 (13.6 %) the S_{SM} coefficient. In most of these studies the similarity coefficients were used without the authors giving any rationale given for their choice, regardless of the species' ploidy level or the degree of genetic recombination or heterozygosity expected of its mating system. We set out to estimate and compare the usefulness of the S_{SM}, S_J, and S_D coefficients in genetic diversity assessment based on RAPD or AFLP data from predominantly self-pollinating species studied by all co-authors. Considering that clustering and ordination results can be influenced by the choice of coefficient (Duarte *et al.* 1999, and literature cited therein) these coefficients need to be better understood.

This study examined 139 *Phaseolus vulgaris* L. (common bean), 67 *Lactuca saligna* L. (least lettuce), and 110 *Triticum dicoccoides* Koern. (wild emmer wheat) accessions, all of which have been used in previous investigations of genetic diversity. All three species are annual and predominantly self-fertilizing plants. The relevant references, as well as details of the origin of accessions, DNA extraction, amplification products and genetic diversity estimates, were reported by Šuštar-Vozlič *et al.* (2006) for *P. vulgaris* and by Kitner *et al.* (2008) for *L. saligna* using AFLP markers, and by Fahima *et al.* (1999) for *T. dicoccoides* using RAPD markers. The problem of analyzing dominant markers in molecular population genetics (Lynch and Milligan 1994) is overcome by the high rate of selfing in all three species under this study.

We analyzed a total of 424 AFLP bands of *P. vulgaris* (Šuštar-Vozlič *et al.* 2006), a total of 490 AFLP bands of *L. saligna* (Kitner *et al.* 2008), and a total of 59 RAPD bands of *T. dicoccoides* (Fahima *et al.* 1999). The basic data structure finally consisted of three binary matrices (0/1), representing the scored markers (monomorphic as well as polymorphic). The binary matrices were subjected to statistical analysis by *NTSYS-pc* version *2.02* (Rohlf 1998). Results were analyzed as a dominant mode of inheritance in a diploid organism. S_{SM}, S_J, and S_D coefficients were used to compute pairwise genetic similarities. The corresponding dendrograms were constructed by the unweighted pair-group method with the arithmetic averages (UPGMA) clustering algorithm. To check the goodness of fit of a cluster analysis with the associated similarity matrix, cophenetic correlation (Sokal and Rohlf 1962) coefficients (CCC, *i.e.*, the correlation of the cophenetic values matrix with the values of the similarity matrix on which the clustering was based) were computed. The clustering was validated by 1000 replicates of the bootstrap analysis using the *WinBoot* computer package (Yap and Nelson 1996). For each species, the degree of congruence between different similarity matrices and between different cophenetic matrices was ascertained by Mantel's matrix correspondence test (Mantel 1967), a randomization procedure that compares correlations of two matrices.

Results and discussion

Genetic similarity values were lowest for S_J (Table 1*A*), as expected, since by definition $S_D \geq S_J$ and $S_{SM} \geq S_J$. S_{SM} values were lower than S_D values, again as expected for the relatively high similarity values found for the three species, since there is a relation between S_{SM} and S_D values depending on the relation between the number of 1-1 matches and the number of 0-0 matches (Kosman and Leonard 2005). Dendrograms were generated by cluster analysis from the genetic similarity matrices created by using the three similarity coefficients (Fig. 1). The differences in the structure of the dendrograms are hardly visible, and accessions' ordering is identical, partly an effect of the relations among the similarity measures ("joint monotonicity"; Sneath and Sokal 1973).

For both *P. vulgaris* and *L. saligna* the three similarity indices were presented by very high ($r > 0.95$) and strongly significant ($P < 0.001$) CCC values (Table 1*B*), indicating a very high level of fit for each cluster

Table 1. Results based on three genetic similarity coefficients: S_{SM} (simple matching), S_D (Dice), and S_J (Jaccard), revealed by AFLP/ RAPD markers from three species: *Phaseolus vulgaris* (AFLP; Šuštar-Vozlič *et al.* 2006), *Lactuca saligna* (AFLP; Kitner *et al.* 2008), *Triticum dicoccoides* (RAPD; Fahima *et al.* 1999). *A* - Average, range and coefficient of variation (CV - standard deviation/ mean). *B* - Mantel's correlation coefficients between cophenetic and similarity matrices (CCC; diagonal), between similarity matrices (below diagonal), and between cophenetic matrices (above diagonal). All correlations were highly significant ($P < 0.001$).

	A				*B*		
		average	range	CV [%]	S_{SM}	S_D	S_J
Phaseolus	S_{SM}	0.890	0.550 - 1.000	7.69	0.981	0.999	0.998
vulgaris	S_D	0.916	0.632 - 1.000	5.90	0.999	0.982	0.995
	S_J	0.849	0.462 - 1.000	10.08	0.998	0.995	0.979
Lactuca	S_{SM}	0.786	0.550 - 1.000	12.50	0.963	0.994	0.996
saligna	S_D	0.819	0.607 - 1.000	10.60	0.990	0.953	0.997
	S_J	0.703	0.436 - 1.000	18.40	0.993	0.997	0.966
Triticum	S_{SM}	0.747	0.525 - 0.983	7.80	0.598	0.847	0.848
dicoccoides	S_D	0.817	0.658 - 0.987	5.30	0.983	0.578	0.999
	S_J	0.693	0.491 - 0.975	8.90	0.982	0.998	0.593

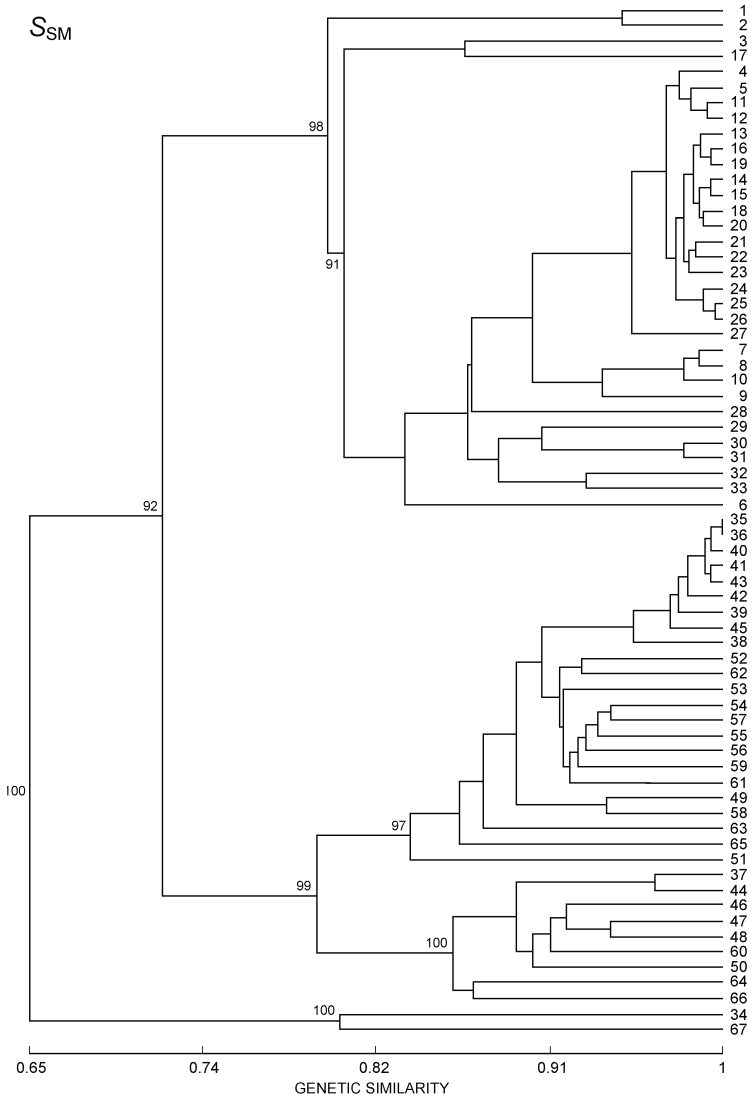

Fig. 1*A*. UPGMA dendrogram based on S_{SM} (simple matching) similarity coefficient among 67 *L. saligna* accessions obtained from 490 AFLP loci (Kitner *et al.* 2008). Numbers on branches are percentage values from bootstrap analysis (1000 replicates).

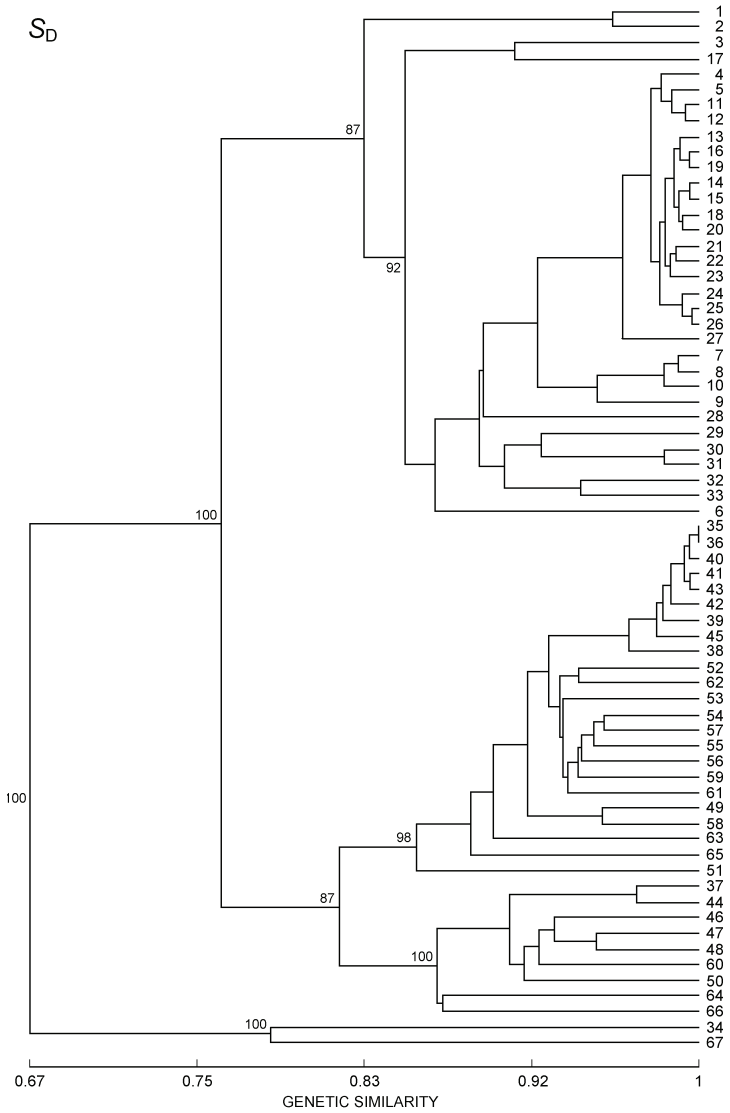

Fig. 1*B*. UPGMA dendrograms based on S_D (Dice) similarity coefficient among 67 *L. saligna* accessions obtained from 490 AFLP loci (Kitner *et al.* 2008). Numbers on branches are percentage values from bootstrap analysis (1000 replicates).

analysis. Significant ($P < 0.001$) but moderate ($0.598 \geq r \geq 0.578$) CCC values were obtained for *T. dicoccoides* by the three similarity indices. Note that only a single pair of accessions from both *P. vulgaris* and *L. saligna* (# 35 and 36, see Fig. 1) generated similarities equal to 1. Thus, this may not be the source of high cophenetic correlations. Very high and strongly significant ($P < 0.001$) correlations were obtained between similarity matrices ($r \geq 0.982$), as well as between cophenetic matrices ($r \geq 0.847$), based on the three genetic similarity coefficients of all three species. The discriminating power was low for *T. dicoccoides* but very high for both *P. vulgaris* and *L. saligna*, as indicated also by bootstrap values of a UPGMA clustering (example of S_{SM} (not presented here): bootstrap values averaged 6, 83, and 97 for the three species, respectively).

Inclusion of negative matches in the calculation of similarity coefficients may raise serious doubts among taxonomists (Romesburg 1984). Sneath and Sokal (1973) discussed this problem in respect of microorganisms and higher organisms for both morphological and metabolic characters. The original argument for excluding 0-0 matches was based on the use of traits where all groups being compared included the 0 state of the trait. By a conservative approach, similarity of DNA fingerprints is generally also defined as the fraction of observed bands that are shared by two individuals (Kosman and Leonard 2005). However, for molecular marker data, inclusion of 0-0 matches seems appropriate when two alleles exist at a locus and one produces a band while the other does not, and both alleles are present in the materials being compared (Dudley 1994). For dominant markers it is generally assumed that each band represents a different bi-allelic locus (Williams *et al.* 1990) and that the alternative to a band at the gel position characteristic of that locus is the absence of a band anywhere in the gel

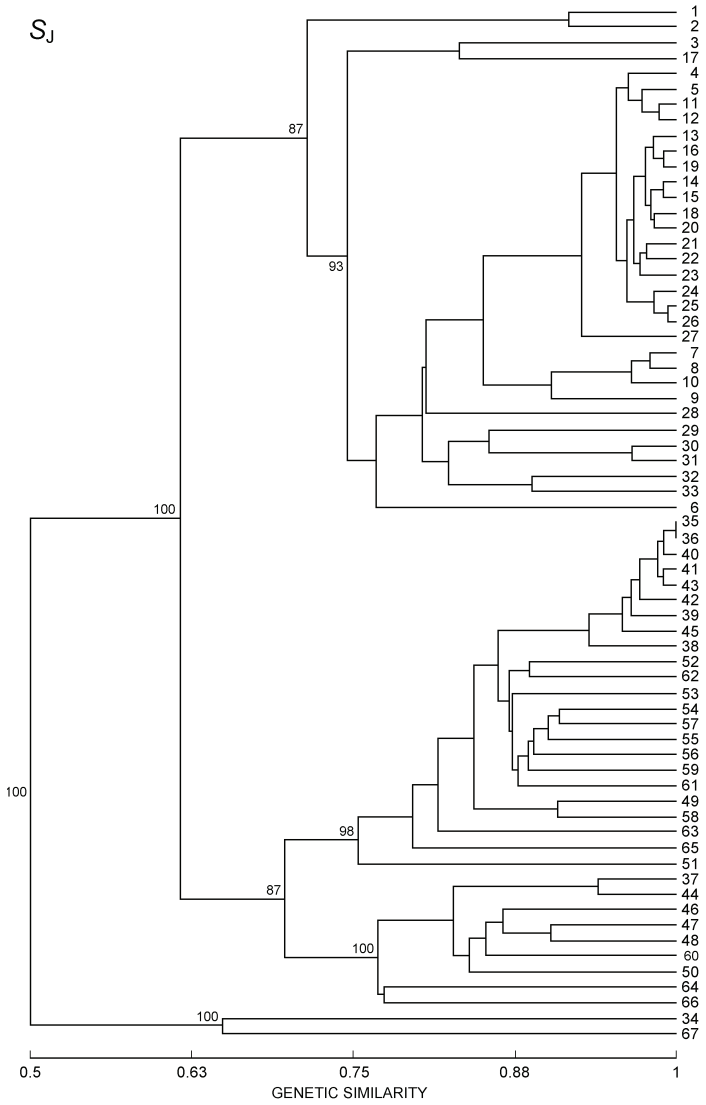

Fig. 1*C*. UPGMA dendrograms based on S_J (Jaccard) similarity coefficients among 67 *L. saligna* accessions obtained from 490 AFLP loci (Kitner *et al.* 2008). Numbers on branches are percentage values from bootstrap analysis (1000 replicates).

(Kosman and Leonard 2005). In addition, for diploid organisms that are primarily inbreeders we would expect a very low level of heterozygosity, so treating those organisms as if they were haploids might be sufficient (Simioniuc *et al.* 2002), *i.e.*, it would be generally possible to determine the exact genetic similarity between two individuals that share a band at the same position. Also, in the case of closely related individuals very low DNA sequence divergence between them is expected. Thus, the S_{SM} coefficient may be the most suitable measure of similarity, as also suggested by Kosman and Leonard (2005) as well as by Zeid *et al.* (2003), who assessed genetic diversity in faba bean lines using AFLP markers. These authors surmised that negative matches revealed a genetic similarity because the absence of a band in two individuals should be due to an identical cause inherited from the same ancestor. This explanation seems most appropriate also for the results of AFLP analysis of *P. vulgaris* and *L. saligna* described in this communication, but it also contains CCC results used as a measure of goodness of fit for the cluster analysis, thereby endorsing the conclusions of existing publications. For these two species, very high CCCs were calculated regardless of the similarity coefficient used. On the other hand, only moderate CCCs were calculated for RAPD analysis of *T. dicoccoides*. A similar value of CCC (r = 0.58) was observed by Vijayan *et al.* (2004) for Indian mulberry varieties using RAPD markers and the S_D similarity coefficient. In establishing genetic relationships in olive using the same marker system and similarity coefficient, Belaj *et al.* (2003) observed CCC of 0.67. Powell *et al.* (1996) reported CCC of 0.794 in *Glycine max* assessed by RAPD markers and the S_{SM} coefficient. The lower CCC observed in all these species by RAPD analysis, including *T. dicoccoides* described here, must be due to the greater genetic similarity of the

accessions studied, as a low cophenetic coefficient is indicative of inconsistency among objects and groupings. In other words, low CCC values suggest a highly complex pattern of multiple resemblances for most individuals, which was difficult for the clustering algorithm to resolve (Barnesky and Lammers 1997).

The RAPD technique often shows weak reproducibility between laboratories for several reasons, described by Schweder *et al.* (1995) and literature cited therein. Also, putatively similar bands originating from RAPDs in different individuals are not necessarily homologous although they may share the same size in base pairs and may therefore lead to wrong results when genetic relationships are calculated. According to Lamboy (1994), optimization of PCR protocols for eliminating artifacts from RAPD data is either not possible or not successful. The three most widely used similarity coefficients with RAPD data, S_D, S_J, and S_{SM}, differ in the amount of bias produced by the level of artifactual bands. Considering several factors, Lamboy (1994) recommended that the S_D coefficient be used routinely for measuring similarities in RAPD data unless specific circumstances or needs dictate the use of the other two coefficients. The same recommendation was made by Duarte *et al.* (1999), who compared similarity coefficients based on RAPD markers in the common bean. In contrast to RAPDs, AFLPs have high capacity to reveal several bands in one lane and have better reproducibility and transferability. But they are technically more demanding than RAPDs, *e.g.*, require sequencing gels. To some extent AFLPs suffer the same drawbacks as RAPDs concerning the data generated: the frequency of erroneous scoring of non-homologous bands of similar mobility may increase for genetically less similar individuals, leading to overestimation of similarity; yet this is less likely to occur for AFLPs because of higher gel resolution.

To sum up, in this study we examined the most adequate similarity coefficient for determining genetic divergence in predominantly self-pollinating species using dominant DNA markers. In the case of closely related individuals, as in our study, the absence of a band in two individuals should be due to an identical cause inherited from the same ancestor. Therefore, we propose the S_{SM} coefficient because negative matches represent a genetic similarity. When RAPDs are employed, inferences about population structure and nucleotide divergence should be made with prudence because the nature of genetic variation uncovered by RAPDs is often unclear.

References

Barnesky, A.L., Lammers, T.G.: Revision of the endemic Asian genus *Peracarpa* (Campanulaceae: Campanuloideae) via numerical phenetics. - Bot. Bull. Acad. sin. **38**: 49-56, 1997.

Belaj, A., Satovic, Z., Cipriani, G., Baldoni, L., Testolin, R., Rallo, L., Trujillo, I.: Comparative study of the discriminating capacity of RAPD, AFLP and SSR markers and of their effectiveness in establishing genetic relationships in olive. - Theor. appl. Genet. **107**: 736-744, 2003.

Cordeiro, A.I., Sanchez-Sevilla, J.F., Alvarez-Tinaut, M.C., Gomez-Jimenez, M.C.: Genetic diversity assessment in Portugal accessions of *Olea europaea* by RAPD markers. - Biol. Plant. **52**: 642-647, 2008.

Dice, L.R.: Measures of the amount of ecologic association between species. - Ecology **26**: 297-302, 1945.

Duarte, J.M., Santos, J.B., Melo, L.C.: Comparison of similarity coefficients based on RAPD markers in the common bean. - Genet. mol. Biol. **22**: 427-432, 1999.

Dudley, J.W.: Comparison of genetic distance estimators using molecular marker data. - In: Analysis of molecular marker data. Pp. 3-7. Amer. Soc. Hort. Sci., Crop Sci. Soc. Amer., Corvallis 1994.

Fahima, T., Sun, G.L., Beharav, A., Krugman, T., Beiles, A., Nevo, E.: RAPD polymorphism of wild emmer wheat populations, *Triticum dicoccoides*, in Israel. - Theor. appl. Genet. **98**: 434-447, 1999.

Gower, J.C.: Measures of similarity, dissimilarity and distances. - In: Kotz, S., Johnson, N.L., Read, C.L. (ed.): Encyclopedia of Statistical Sciences. Vol. 5. - Wiley, New York 1985.

Jaccard, P.: Nouvelles recherches sur la distribution florale. - Bull. Soc. Vaud. Sci. natur. **44**: 223-270, 1908.

Kitner, M., Lebeda, A., Doležalová, I., Maras, M., Křístková, E., Beharav, A., Nevo, E., Pavlíček, T., Meglič, V.: AFLP analysis of *Lactuca saligna* germplasm collections from four European and three Middle East countries. - Isr. J. Plant Sci. **56**: 185-193, 2008.

Kosman, E., Leonard, K.J.: Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. - Mol. Ecol. **14**: 415-424, 2005.

Lamboy, W.F.: Computing genetic similarity coefficients from RAPD data: the effects of PCR artifacts. - PCR Methods Appl. **4**: 31-37, 1994.

Lynch, M., Milligan, B.: Analysis of population genetic structure with RAPD markers. - Mol. Ecol. **3**: 91-99, 1994.

Mantel, N.A.: The detection of disease clustering and a generalized regression approach. - Cancer Res. **27**: 209-220, 1967.

Nei, M., Li, W.H.: Mathematical models for studying genetic variation in terms of restriction endonucleases. - Proc. nat. Acad. Sci. USA **76**: 5269-5273, 1979.

Powell, W., Morgante, M., Andre, C., Hanafey, M., Vogel, J., Tingey, S., Rafalski, A.: The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. - Mol. Breed. **2**: 225-238, 1996.

Reif, J.C., Melchinger, A.E., Frisch, M.: Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. - Crop Sci. **45**: 1-7, 2005.

Rohlf, F.J.: NTSYS-pc. Numerical Taxonomy and Multivariate Analysis System. - Applied Biostatistics, New York 1998.

Romesburg, H.C.: Cluster Analysis for Researchers. - Krieger Publ. Comp., Malabar 1984.

Schweder, M.E., Shatters, R.G., West, S.H., Smith, R.L.: Effect of transition interval between melting and annealing temperatures on RAPD analyses. - Biotechniques **38**: 40-42,

1995.

Simioniuc, D., Uptmoor, R., Friedt, W., Ordon, F.: Genetic diversity and relationships among pea cultivars revealed by RAPDs and AFLPs. - Plant Breed. **121**: 429-435, 2002.

Sneath, P.H.A., Sokal, R.R.: Numerical Taxonomy. The Principles and Practice of Numerical Classification. - W.H. Freeman and Company, San Francisco 1973.

Sokal, R.R., Michener, C.D.: A statistical method for evaluating systematic relationships. - Univ. Kansas Sci. Bull. **38**: 1409-1438, 1958.

Sokal, R.R., Rohlf, F.J.: The comparison of dendrograms by objective methods. - Taxon **11**: 33-40, 1962.

Šuštar-Vozlič, J., Maras, M., Javornik, B., Meglič, V.: Genetic diversity and origin of Slovene common bean (*Phaseolus vulgaris* L.) germplasm as revealed by AFLP markers and phaseolin analysis. - J. amer. Soc. hort. Sci. **131**: 242-249, 2006.

Vijayan, K., Awasthi, A.K., Srivastava, P.P., Saratchandra, B.: Genetic analysis of Indian mulberry varieties through molecular markers. - Hereditas **141**: 8-14, 2004.

Weising, K., Nybom, H., Wolff, K., Meyer, W.: DNA Fingerprinting in Plants and Fungi. - CRC Press, Boca Raton 1995.

Williams, J.G.K., Kubelik, A.R., Llivak, K.J., Rafalski, J.A., Tingey, S.V.: DNA polymorphism amplified by arbitrary primers are useful as genetic markers. - Nucl. Acids Res. **18**: 6531-6535, 1990.

Yap, I.V., Nelson, R.J.: WinBoot: A Program for Performing Bootstrap Analysis of Binary Data to Determine the Confidence Limits of UPGMA-based Dendograms. - International Rice Research Institute, Manila 1996.

Zeid, M., Schon, C.C., Link, W.: Genetic diversity in recent elite faba bean lines using AFLP markers. - Theor. appl. Genet. **107**: 1304-1314, 2003.